Adaptive Proposer for Ultimatum Game

František Hůla, Marko Ruman and Miroslav Kárný*

Department of Adaptive Systems,
Institute of Information Theory and Automation, Czech Academy of Sciences
POB 18, 182 08 Prague 8, the Czech Republic
{hula.frantisek,marko.ruman}@gmail.com, school@utia.cas.cz
http://www.utia.cz/AS

Abstract. Ultimate Game serves for extensive studies of various aspects of human decision making. The current paper contribute to them by designing proposer optimising its policy using Markov-decision-process (MDP) framework combined with recursive Bayesian learning of responder's model. Its foreseen use: i) standardises experimental conditions for studying rationality and emotion-influenced decision making of human responders; ii) replaces the classical game-theoretical design of the players' policies by an adaptive MDP, which is more realistic with respect to the knowledge available to individual players and decreases player's deliberation effort; iii) reveals the need for approximate learning and dynamic programming inevitable for coping with the curse of dimensionality; iv) demonstrates the influence of the fairness attitude of the proposer on the game course; v) prepares the test case for inspecting exploration-exploitation dichotomy.

Keywords: Games; Markov decision process; Bayesian learning

1 Introduction

Since ancient times people trade with each other. Modern man cannot imagine life without exchange of goods, services, information etc. It is something like the cornerstone of our civilization. This human activity divides people to proposers of some merit and responders who either accept or refuse it. Both of them can often bargain but a price tag in the store represents a sort of ultimatum: if we buy the product we agree with the seller's price without any direct negotiation. This motivates investigations of human behaviour connected with the bargaining and trading. They concern economical, game-theoretical, social, cultural and emotional aspects and they often use standardised "laboratory" variants of the discussed interaction. Ultimatum Game (UG) is a prominent test case [18].

UG considers a fixed number of rounds of the two-player game. A fixed amount of money is split in each round. The proposer offers a part of this amount and the responder either accepts the offer and money are split accordingly or refuses it and both get nothing. Seemingly, the game should have a definite course:

* This research was supported by Grant Agency of the Czech Republic, No 13-13502S.

the proposer offers the smallest possible positive amount and the responder accepts it. No such behaviour is observed in reality as people judge the game not only according to monetary profit. Typically, they try to earn at least as their opponent, they care about self-fairness [8,20] influenced by culture, sex, etc. [5]. An important influence of emotions are also studied [7].

While an appropriate model of self-fairness leads to surprisingly accurate predictions of responders' behaviours [9], to get a statistically significant quantification of emotional influences has been found quite hard. The hypothesis that an actively optimising proposer could make this influence more pronounced has led to the design of the standardised active proposer described here. Theory of Markov decision processes (MDP) [17] was selected as the basis of such a design. This choice avoiding the standard game-theoretical formulation [22] is motivated by the inherent trap of the dimensionality curse [3] of the Bayesian games [10].

The use of MDP supposes knowledge of responder's model, which is in realistic scenarios unknown and, moreover, it is very individual in the targeted emotion-oriented studies. This calls for a combination of MDP with a permanent learning of this model, i.e. for adaptive MDP. The inherent small amount of available data singles out recursive Bayesian learning in the closed decision loop as the (only) appropriate methodology [16]. Even then, approximations of recursive learning like [11] and dynamic programming [21] are needed. This text makes just the preparatory steps towards the complete solution. It recalls MDP, the dynamic programming as the optimisation tool and the recursive Bayesian learning, Section 2. UG is formulated in MDP terms for various types of proposers, Section 3. A numerical illustration is in Section 4. Section 5 adds remarks.

2 Mathematical Background

This section is based on [3,9,16,17]. It introduces the adopted notions, recalls the used mathematical tools, and makes the paper relatively self-containing.

2.1 General Formulation and Solution of Markov Decision Process

The considered system consists of *decision maker* (DM), and *responder*. They interact in discrete time (*decision epochs*) $t \in \mathbf{T} = \{1, 2, \dots, |\mathbf{t}|\}$, $|\mathbf{t}| < \infty$. DM chooses a discrete-valued *action* (an irreversible decision) $a_t \in \mathbf{A} = \{1, 2, \dots, |\mathbf{a}|\}$, $|\mathbf{a}| < \infty$, in each epoch $t \in \mathbf{T}$. Consequently, the closed decision loop transits from a discrete-valued *state* $s_{t-1} \in \mathbf{S} = \{1, 2, \dots, |\mathbf{s}|\}$, $|\mathbf{s}| < \infty$ to the state $s_t \in \mathbf{S}$. The use of *regression pair* $\psi_t = (a_t, s_{t-1}) \in \mathbf{\Psi} = (\mathbf{A}, \mathbf{S})$, $t \in \mathbf{T}$, simplifies the presentation. With it, the random transition is described by *transition probabilities*[1] $\big(p(s_t|\psi_t)\big)_{t\in\mathbf{T}} \in \mathbf{P} = \Big\{p(s_t|\psi_t) \geq 0 \,\Big|\, \sum_{s_t\in\mathbf{S}} p(s_t|\psi_t) = 1, \ \forall \psi_t \in \mathbf{\Psi}\Big\}$. After the transition, the DM receives a real-valued *reward* $r(s_t, \psi_t) \in \mathbf{R} = \Big\{r(s_t, \psi_t) \,\Big|\, s_t \in \mathbf{S}, \, \psi_t \in \mathbf{\Psi}, \, t \in \mathbf{T}\Big\}$. The DM cannot use the state s_t for choosing the action $a_t \in \mathbf{A}$ and thus it can at most maximise *aggregate expected reward*

[1] All functions with time-dependent arguments generally depend on time.

$$\begin{aligned}\sum_{t\in\mathbf{T}} \mathsf{E}[r(s_t,\psi_t)] &= \sum_{t\in\mathbf{T}} \sum_{s_t\in\mathbf{S},\psi_t\in\boldsymbol{\Psi}} r(s_t,\psi_t)p(s_t,\psi_t) \qquad (1)\\ &= \sum_{t\in\mathbf{T}} \sum_{\substack{s_t\in\mathbf{S},a_t\in\mathbf{A}\\ s_{t-1}\in\mathbf{S}}} r(s_t,\psi_t)p(s_t|\psi_t)p(a_t|s_{t-1})p(s_{t-1}).\end{aligned}$$

The last equality in (1) follows from the chain rule [17]. It expresses the probability $p(s_t,\psi_t)$ as the product of the given *transition probability* $p(s_t|\psi_t)\in\mathbf{P}$, of the optional *decision rules* $(p(a_t|s_{t-1}))_{t\in\mathbf{T}}\in\boldsymbol{\Pi}=\Big\{p(a_t|s_{t-1})\Big|s_{t-1}\in\mathbf{S},a_t\in\mathbf{A}\Big\}$ forming the *decision policy* and of the *state probability* $p(s_t)\in\mathbf{PS}$, where

$$\mathbf{PS}=\Big\{p(s_t)\,\Big|\,p(s_t)=\sum_{\psi_t\in\boldsymbol{\Psi}} p(s_t|\psi_t)p_t(a_t|s_{t-1})p_t(s_{t-1}),\ s_t\in\mathbf{S},\ t\in\mathbf{T}\Big\}. \qquad (2)$$

The state probability $p(s_t)$ is influenced by the "policy prefix" $\big(p(a_\tau|s_{\tau-1})\big)_{\tau\leq t}$ and the probability $p(s_0)$ of the initial state $s_0\in\mathbf{S}$. Often, $p(s_0)=\delta(s_0,\tilde{s}_0)$ with *Kronecker* δ equal to 1 for equal arguments and 0 otherwise. It concentrates $p(s_0)$ on a given $\tilde{s}_0\in\mathbf{S}$.

Thus, the optimising DM maximises the aggregate expected reward (1) over *decision policies* $\boldsymbol{\Pi}$. For known $\mathbf{R}$, $\mathbf{P}$ and $\tilde{s}_0$, DM takes as the *optimal policy*

$$(p_{opt}(a_t|s_{t-1}))_{t\in\mathbf{T}}\in \operatorname{Arg}\max_{(p(a_t|s_{t-1}))_{t\in\mathbf{T}}\in\boldsymbol{\Pi}} \sum_{t\in\mathbf{T}} \mathsf{E}[r(s_t,\psi_t)]. \qquad (3)$$

Definition 1 (Optimal MDP). *The given 7-tuple* $\{\mathbf{T},\mathbf{A},\mathbf{S},\mathbf{PS},\mathbf{P},\mathbf{R},\boldsymbol{\Pi}\}$ *together with the maximisation (3) is referred as Markov decision process (MDP).*

Theorem 1 (Dynamic Programming, proof e.g. in [17]). *The policy* $(p_{opt}(a_t|s_{t-1}))_{t\in\mathbf{T}}\in\boldsymbol{\Pi}$ *maximising the aggregate expected reward (3) consists of the deterministic decision rules* $p_{opt}(a_t|s_{t-1})=\delta(a_t,a_t^\star(s_{t-1}))$*, where*

$$a_t^\star(s_{t-1})\in\operatorname{Arg}\max_{a\in\mathbf{A}}\mathsf{E}[r(s_t,a,s_{t-1})+\varphi_t(s_t)|a,s_{t-1}]\ \textit{with}\ \text{value function}$$

$$\varphi_t(s_t)=\sum_{s_{t+1}\in\mathbf{S}}[r(s_{t+1},a_{t+1}^\star(s_t),s_t)+\varphi_{t+1}(s_{t+1})]p(s_{t+1}|a_{t+1}^\star(s_t),s_t).$$

The backward recursion starts with $\varphi_{|\mathbf{t}|}(s_{|\mathbf{t}|})=0,\forall s_{|\mathbf{t}|}\in\mathbf{S}$.

2.2 Bayesian Learning of Transition Probabilities

The unrealistic assumption that the transition probabilities from the set $\mathbf{P}$ are given, see Definition 1, is removed via Bayesian recursive learning [16] recalled here. It relates the state $s_t\in\mathbf{S}$ to the action $a_t\in\mathbf{A}$ and observed states $s_\tau\in\mathbf{S}$, $\tau<t$, by transition probability parameterised by its unknown values Θ

$$p(s_t|a_t,\ldots,a_1,s_{t-1},\ldots,s_0,\Theta)=p(s_t|\psi_t,\Theta)=\prod_{s\in\mathbf{S}}\prod_{\psi\in\boldsymbol{\Psi}}\Theta_{s|\psi}^{\delta(s,s_t)\delta(\psi,\psi_t)},\ \text{where}$$

$$\Theta\in\boldsymbol{\Theta}=\Big\{\Theta_{s|\psi}\geq 0\,\Big|\,s\in\mathbf{S},\psi\in\boldsymbol{\Psi},\ \sum_{s\in\mathbf{S}}\Theta_{s|\psi}=1,\ \forall\psi\in\boldsymbol{\Psi}\Big\}. \qquad (4)$$

It provides the transition probabilities as predictors

$$p(s_t|\psi_t,\dots,\psi_1) = \int_{\mathbf{\Theta}} p(s_t|\psi_t,\Theta)p(\Theta|s_{t-1},\psi_{t-1},\dots,\psi_1)\mathrm{d}\Theta, \tag{5}$$

where the posterior probability density $p(\Theta|s_{t-1},\psi_{t-1},\dots,\psi_1)$ has the support $\mathbf{\Theta}$ and is given by the observed condition $s_{t-1},\psi_{t-1},\dots,\psi_1$. Bayes' rule [4,16] evolves it

$$p(\Theta|s_t,\psi_t,\dots,\psi_1) = \frac{p(s_t|\psi_t,\Theta)p(\Theta|s_{t-1},\psi_{t-1},\dots,\psi_1)}{p(s_t|a_t,\psi_{t-1},\dots,\psi_1)}. \tag{6}$$

An optional prior probability density $p(\Theta) = p(\Theta|\psi_1,\psi_0)$ initiates (6).

Importantly, the learning (5), (6) is valid for any policy for which the parameter $\Theta \in \mathbf{\Theta}$ is unknown, i.e. which meets *natural conditions of control* [16]

$$p(a_t|\psi_{t-1},\dots,\psi_1,\Theta) = p(a_t|\psi_{t-1},\dots,\psi_1). \tag{7}$$

The learning is correct in loops closed by any (say human) policy meeting (7).

The product forms of the model (4) and of Bayes' rule (6) imply Dirichlet's form of the posterior probability density, [12], which uses Euler's gamma $\mathsf{\Gamma}$ [1],

$$p(\Theta|\psi_t,\dots,\psi_1) = p(\Theta|V_t) =, \prod_{\psi\in\mathbf{\Psi}} \mathsf{\Gamma}\Big(\sum_{\tilde{s}\in\mathbf{S}} V_{t;\tilde{s}|\psi}\Big)\frac{\prod_{s\in\mathbf{S}}\Theta_{s|\psi}^{V_{t;s|\psi}-1}}{\mathsf{\Gamma}(V_{t;s|\psi})}, \quad \text{where}$$
$$V_{t;s|\psi} = V_{t-1;s|\psi} + \delta(s,s_t)\delta(\psi,\psi_t), \text{ form } \textit{occurence array}, \ s\in\mathbf{S},\psi\in\mathbf{\Psi}. \tag{8}$$

The initial occurrence array $V_0 = (V_{0;s|\psi} > 0)_{s\in\mathbf{S},\psi\in\mathbf{\Psi}}$ describes the used conjugated (Dirichlet form preserving) prior probability density $p(\Theta)$. The gained predictive probability resembles the frequentist estimate $\hat{\Theta} \in \mathbf{\Theta}$ of $\Theta \in \mathbf{\Theta}$

$$p(s_t = s|\psi_t = \psi, V_{t-1}) = \frac{V_{t-1;s|\psi}}{\sum_{\tilde{s}\in\mathbf{S}} V_{t-1;\tilde{s}|\psi}} = \hat{\Theta}_{t-1;s|\psi}, \ s\in\mathbf{S}, \ \psi\in\mathbf{\Psi}. \tag{9}$$

3 Ultimatum Game as Adaptive MDP

According to UG rules, $|\mathbf{t}|$ (tens) rounds are played. Possible actions $a_t \in \mathbf{A}$ of the proposer P (DM supported here) in the round $t \in \mathbf{T}$ are the offered splits of $q = |\mathbf{a}| + 1$ (often monetary) units. The responder R generates the *observed response* $o_t \in \mathbf{O} = \{1,2\} = \{\text{reject the offer}, \text{accept the offer}\}$. The profits of the proposer $Z_{t;P}$ and responder $Z_{t;R}$ accumulated after tth round are

$$Z_{t;P} = \sum_{\tau=1}^{t}(q-a_\tau)(o_\tau - 1) \in \mathbf{Z}_{t;P}, \ Z_{t;R} = \sum_{\tau=1}^{t} a_\tau(o_\tau - 1) \in \mathbf{Z}_{t;R}. \tag{10}$$

The profits (10) determine the observable (non-minimal) state s_t of the game

$$s_t = (Z_{t;P}, Z_{t;R}) \in \mathbf{S} = (\mathbf{Z}_{t;P}, \mathbf{Z}_{t;R}), \ t \in \mathbf{T}. \tag{11}$$

It has a finite amount of values and starts with zero profits $s_0 = (0, 0)$.

Altogether, UG rules directly specify sets of epochs $\mathbf{T}$, actions $\mathbf{A}$, and state probabilities $\mathbf{PS}$, cf. (2), in the 7-tuple delimiting MDP, see Definition 1. The peculiarities of the use of adaptive MDP by the proposer thus reduce to those connected with transition probabilities $\mathbf{P}$, rewards $\mathbf{R}$ and with the curse of dimensionality connected with the policy space $\mathbf{\Pi}$. They are discussed below.

3.1 Transition Probabilities

Bayesian learning, Section 2.2, formally provides the needed transition probabilities in $\mathbf{P}$ under acceptable assumption that the responder does not vary them abruptly during the game course. The lack of learning data, consequence of the dimensionality curse [3], is, however, serious obstacle. Indeed, the sufficient statistics V_t (8) has $|\mathbf{o}| \times |\mathbf{\Psi}| = |\mathbf{o}| \times |\mathbf{a}| \times |\mathbf{s}|$ entries. In a typical case $|\mathbf{o}| = 2$, $|\mathbf{a}| = 9$ and reduced $|\mathbf{s}| = 10$, it needs 180 values to be populated by data, which requires unrealistic hundreds' game rounds. The ways out are as follows.

Reduction of the State Space: The size of V is determined by the richness of the state space $\mathbf{S}$. The UG rules imply that the two-dimensional s_t (11) stays in $(s_{t-1}, s_{t-1} + [a_t, q - a_t])$, i.e. many transitions are impossible. The above example, respecting this fact, indicates the need for additional countermeasures.

Use of Population Based Priors: It is possible to obtain a reliable description of responders' population and convert it into the prior occurrence array $V_{0;s|\psi} = v_{0;\psi}\hat{\Theta}_{0;s|\psi}$, $s \in \mathbf{S}$, $\psi \in \mathbf{\Psi}$ (9). V_0 is modified by at most $|\mathbf{t}|$ data records specific for the individual responder in the individual game. Thus, the choice of the *prior weight* $v_{0;\psi} > 0$ is critical. Due to data sparsity, a few observations of a specific s, ψ within tens of rounds are expected. Thus, $v_{0;\psi} \leq |\mathbf{t}|$ is recommendable.

Assuming ψ-independent prior weight $v_0 = v_{0;\psi}$, its hierarchical Bayesian learning [4] becomes feasible. Hypotheses $h:$ proper $v_0 = v_{0;h} =$ a value in (0,1), $h \in \mathbf{H} = \{1, 2, \ldots, |\mathbf{h}|\}$ with a small $|\mathbf{h}|$ are formulated. For each h, the predictor (9), becomes h dependent $p(s_t = s|\psi_t = \psi, V_{t-1;h})$ via h-dependent array $V_{0;h} = v_{0;h}\hat{\Theta}$. Then, Bayes' rule is directly applicable. The h-independent predictor (transition probability) becomes mixture of predictors within respective hypotheses with weights being their posterior probabilities, see [4,16].

This Bayesian averaging is of a direct relevance for the motivating studies of influence of emotions on decision making. It suffices to collect descriptions of sub-populations differing by observed or stimulated emotional states and compare hypotheses about suitability of the transition probabilities learnt with prior parameter probability densities reflecting these sub-populations.

Choice of the Model Structure: The above Bayesian treatment of the finite amount of compound hypotheses can also serve for the desirable reduction of the parametric-model structure. For instance, experimental evidence strongly indicates, e.g. [9], that the proposer action decisively influences responder's response. Thus, the hypothesis that $p(s_t|a_t, s_{t-1}) = p(s_t - s_{t-1}|a_t)$ can be and should be compared to the general form of $p(s_t|a_t, s_{t-1})$. It fits to attempts to use more parsimonious parametrisation like special mixtures in [11] are.

3.2 Rewards

The reward $r(s_t, \psi_t) \in \mathbf{R}$, used in the design of the optimal policy (3) reflects attitude of the proposer to inter-relation of its profit and responder's profit. The sole action values play no role unless they are connected with DM's deliberation effort as in [19]. Then, rewards studied in [9] from responder's view-point, are worth considering.

Economic proposer: It is interested in its own profit only, paying no attention to co-player. Its reward is $r(s_t, \psi_t) = Z_{t;P} - Z_{t-1;P}$. It is taken as economically rational DM but almost nobody acts in the way optimal for this reward.

Self-interested proposer: It partially maximises its profit but also watches the responder's profit not to let the responder to win too much. Such attitude was modelled by $r(s_t, \psi_t) = wZ_{t;P} - (1-w)Z_{t;R}$, with the weight $w \in [0, 1]$ controlling self-fairness level. This reward quite successfully models human responders when the weight w is recursively personalised [14,9]. The weight is conjectured to depend on the player's personality and emotions. The preliminary results confirm this [2], but statistically convincing results are unavailable. The adaptive proposer's discussed here is expected to help in this respect.

Fair responder: It jointly maximises profits of both players by using $r(s_t, \psi_t) = wZ_{t;P} - (1-w)\text{abs}(Z_{t;P} - Z_{t;R})$, with the weight $w \in [0, 1]$ balancing own profit with the difference of both profits. No human responder's policy has indicated adoption of such a reward [9]. But the performed experiments limited to greedy (one-stage-ahead) optimisation and the adoption of proposer's view point make us to inspect this variant.

3.3 Policy

The last item to be commented is the set of policies $\mathbf{\Pi}$ within which (approximate) optimum is searched. The described adaptive design extremely increases the extent of the state space as the sufficient statistics V_t (8) is a part of the (information) state. It is obvious as the value function in dynamic programming, Proposition 1, depends on it. This reflects that the selected actions influence not only rewards but also future statistics. The optimal policy is to properly balance the dual – exploitation and exploration – features optimal actions [6]. At present, we are giving it up and use certainty equivalent policies, which perform dynamic programming with the newest parameter estimate $\hat{\Theta}$ taken as known transition probabilities. If need be, the known divergence danger [15] can be overcome by randomising the proposed policy. Foreseen ways are out of our scope.

4 Illustrative Experiments

The limited extent of the paper prevents us to report properly on performed experiments. The illustrative one split $q = |\mathbf{a}| + 1 = 10$, in each of $|\mathbf{t}| = 10$ rounds. The self-fair proposer used the reward $0.5Z_{t;P} - 0.5Z_{t;P}$ and the responder used a fixed randomised decision rule given by the probability $p(o_t = 2 =$

$accept|\psi_t) = p(o_t = 2 = accept|a_t)$, $a_t \in \mathbf{A} = \{1, \ldots, 9\}$. The proposer assumed the same structure but the values $\Theta_{o=2|a}$, $a \in \mathbf{A}$, were recursively estimated, see Section 2.2, and used in designing certainty-equivalent strategy found by dynamic programming, Theorem 1. Samples of experiments running with different weights of the prior estimate v_0 are in Figure 1, where also the used responder's description is visible.

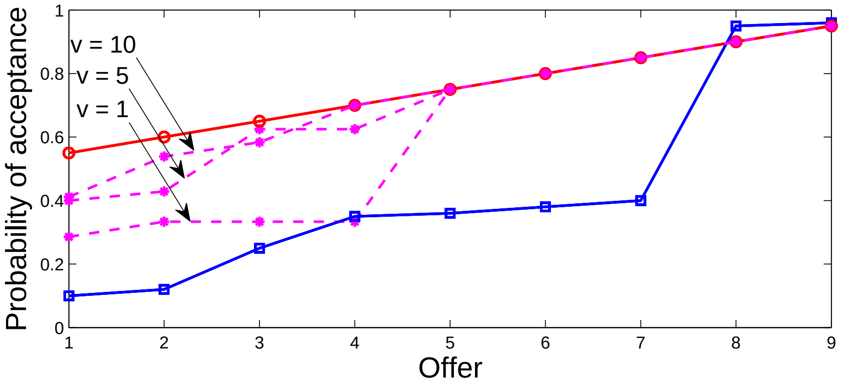

Fig. 1. Final estimates of acceptance probability for possible offers (actions $a \in \mathbf{A}$) are displayed for several weights v= v_0 of the prior occurrence array $V_0 = v_0\hat{\Theta}_0$ and are marked by dots connected with violet dashed lines for distinguishing of each game. The simulated values $\Theta_{o=2|a}, a \in \mathbf{A}$ are marked by squares and the prior values $\hat{\Theta}_{o=2|a}, a \in \mathbf{A}$ are marked by circles. Both of them are connected with blue and red line respectively.

The results are just illustrative and correspond with the expected behaviour: too high weight v_0 makes a significant correction of the prior estimate by a few available data impossible.

5 Conclusion

The paper contributes to a wider research oriented towards influence of personal characteristics, emotional states and available deliberation resources on decision making. Unreported experiments with the proposed optimising adaptive proposer indicate that it can serve to this purpose. On its own, it reveals general problems related to curse of dimensionality and offers test-bed for a further development of techniques fighting with it. Addressing of exploitation-exploration dichotomy is the nearest foreseen problem. In this respect, different types of proposers behaved differently: the economic and self-fair ones, unlike the fair one, exhibited tendency to select a narrow range of actions and as such they are more prone to divergence from optimum. The use of randomised strategies resulting from fully probabilistic design of decision policies [13] seems to be the proper direction.

References

1. M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions.* Dover Publications, New York, 1972.
2. G. Avanesyan. Decision making in ultimatum game. Master's thesis, University of Economics, Prague, 2014.
3. R.E. Bellman. *Adaptive Control Processes.* Princeton U. Press, NJ, 1961.
4. J.O. Berger. *Statistical Decision Theory & Bayesian Analysis.* Springer, NY, 1985.
5. R. Boyd. Cross-cultural Ultimatum Game Research Group - Rob Boyd, Joe Henrich problem. unpublished article.
6. A.A. Feldbaum. Theory of dual control. *Autom. Remote Control*, 21(9), 1960.
7. M. Fiori, A. Lintas, S. Mesrobian, and A.E.P. Villa. Effect of emotion and personality on deviation from purely rational decision-making. In T.V. Guy, M. Kárný, and D.H. Wolpert, editors, *Decision Making & Imperfection.* Springer, Berlin, 2013.
8. W. Güth. On ultimatum bargaining experiments: A personal review. *Journal of Economic Behavior & Organization*, 27(3):329–344, 1995.
9. T.V. Guy, M. Kárný, A. Lintas, and A.E.P. Villa. Theoretical models of decision-making in the ultimatum game: Fairness vs. reason. In R. Wang and Xiaochuan X. Pan, editors, *Advances in Cognitive Neurodynamics (V), Proceedings of the Fifth International Conference on Cognitive Neurodynamics.* Springer, 2015.
10. J.C. Harsanyi. Games with incomplete information played by Bayesian players, I–III. *Management Science*, 50(12), 2004. Supplement.
11. M. Kárný. Recursive estimation of high-order Markov chains: Approximation by finite mixtures. *Information Sciences*, 326:188 – 201, 2016.
12. M. Kárný, J. Böhm, T. V. Guy, L. Jirsa, I. Nagy, P. Nedoma, and L. Tesař. *Optimized Bayesian Dynamic Advising: Theory and Algorithms.* Springer, 2006.
13. M. Kárný and T. Kroupa. Axiomatisation of fully probabilistic design. *Information Sciences*, 186(1):105–113, 2012.
14. Z. Knejflová, G. Avanesyan, T.V. Guy, and M. Kárný. What lies beneath players' non-rationality in ultimatum game? In T.V. Guy and M. Kárný, editors, *Proc. of the 3rd Int. Workshop on Scalable Decision Making, ECML/PKDD 2013.* 2013.
15. P.R. Kumar. A survey on some results in stochastic adaptive control. *SIAM J. Control and Applications*, 23:399–409, 1985.
16. V. Peterka. Bayesian approach to system identification. In *In Trends and Progress in System Identification, P. Eykhoff, Ed*, pages 239–304. Pergamon Press, 1981.
17. M.L. Puterman. *Markov Decision Processes.* Wiley, 1994.
18. A. Rubinstein. Perfect equilibrium in a bargaining model. *Econometrica*, 50(1):97–109, 1982.
19. M. Ruman, F. Hůla, M. Kárný, and T.V. Guy. Deliberation-aware responder in multi-proposer ultimatum game. In *Proceedings of ICANN 2016.* 2016.
20. A.G. Sanfey, J.K. Rilling, J.A. Aronson, L.E. Nystrom, and J.D. Cohen. The neural basis of economic decision-making in the ultimatum game. *Science*, 300(5626):1755–1758, 2003.
21. J. Si, A.G. Barto, W.B. Powell, and D. Wunsch, editors. *Handbook of Learning and Approximate Dynamic Programming*, Danvers, May 2004. Wiley-IEEE Press.
22. J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, New York, London, 1944.